DOCUMENT RESUME

ED 148 724 SP 011 431

AUTHOR Patton, Michael Q.
TITLE Reflections on Evaluating Teacher Centers: The View from an Iron Cow.
PUB DATE 77
NOTE 30p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Data Collection; *Evaluation; Evaluation Methods; *Fundamental Concepts; *Information Theory; *Program Development; *Questioning Techniques; Teacher Centers
IDENTIFIERS *Analysis Objectives

ABSTRACT

This paper focuses on evaluation questions for teacher centers, with the emphasis being that the question is as important as the answer. The questioning may be ordered in nine points: (1) For whom are the questions and answers useful? (2) What effects are desired? How should teachers change after the program? (3) How will this change be observed/described/measured? (4) What are the success/failure criteria for change? At what level/degree are effects expected? (5) What activities and processes are to be used in effecting change? (6) How will the implementation of selected processes be observed/described/selected? (7) What are the success/failure criteria for process/activity implementations? (8) What is the relationship between processes and change? Is there a relationship? (9) What do the answers to the preceeding questions mean? What has been learned that gives a direction for action? When these questions have been answered, and a plan formulated to reflect those answers, evaluation may profitably begin. (MJB)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Michael Q. Patton
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND THE ERIC SYSTEM CONTRACTORS."

ED148724

REFLECTIONS ON EVALUATING TEACHER CENTERS: THE VIEW FROM AN IRON COW

by Michael Q. Patton
Minnesota Center for Social Research
University of Minnesota

"What is the sound of one hand clapping?" This question or riddle was first posed by the Japanese Zen Master Hakuin (1686-1769) as a means of facilitating enlightenment. "The disciple, given a Koan [riddle] to see through, was encouraged to put his whole strength into the singleminded search for its solution, to be 'like a thirsty rat seeking for water...,' to carry the problem with him everywhere, until suddenly, if he were successful, the solution came."[1]

The Koan was a technique originated by the Zen masters to help them shake their students out of routine ways of thinking and acting, to open up new possibilities, and to help the individual student realize his full potential. In many ways the master teacher in a teacher center is engaged in these same processes, helping classroom teachers investigate for themselves different ways of thinking, acting, teaching, and learning. From my perspective, evaluation is yet another way of serving this same function for program staff. The evaluation assists program staff to stand outside themselves and look at what they're doing; evaluations can shake staff out of routine ways of doing things, open up new possibilities, and help staff members realize their full potential.

The starting point in the evaluative process is asking the right questions. A basic tenet of the Koan educational method is that the question is as important as the answer. This paper will focus on evaluation questions for teacher centers. The answers to these questions may be as simple, or as complex, as the answer

SP011 431

---

[1]Yoel Hoffman, The Sound of One Hand (New York, 1975), p. 22.

to any given Zen Koan.

A Flower in Bloom

A monk asked Master Ummon, "What is the pure body of truth?"

Master Ummon said, "A flower in bloom."

Monk: "'A flower in bloom' - what's it mean?"

Master: "Maggots in the shit hole, pus of leprosy, scab over a boil."

Without monk apprentices asking "What's it mean?" there would have been no Zen master in this eighteenth century Oriental version of a teacher center interaction. "What's it mean?" may be a philosophical, religious, or epistemological question. It can also be the very concrete, practical question of researchers or program staff pouring over pages of statistical tables and reams of computer print-out generated by an evaluation study.

Consider a real example from some of the data collected during the 1973-74 evaluation of the pilot teacher centers conducted for the Office of Education. Among educators interviewed in Rhode Island

> ....27 percent of the total sample indicated that they had participated in from one to eight different inservice training sessions sponsored by the RITC. This included 78 percent of the superintendents, 39 percent of the administrators, 1 percent of the secondary school teachers and 22 percent of the elementary school teachers sampled. About 40 percent of the respondents reported that they received graduate credit for their time. The participants generally said that the training met their needs, but 60 percent indicated that this need still existed after the training session.[2]

What's it mean? The answer to that question would seem to depend on who is interpreting the data. For some the data may indicate "a flower in bloom;" for others it may mean "maggots in the shit hole;" and for Master Ummon it might mean both.

---

[2]Robert Covert, John Radzikoski, and Jane Siegel, Evaluating the Four Teacher Center Pilots: The Second Annual Report, Evaluation Research Center, University of Virginia (Charlottesville, Va.: 1974), p. 22.

Sharon Feiman, in commenting on the report cited above, takes issue with the conclusion that "inservice training was made more relevant and current" based only on evidence that "training was offered in a wide variety of innovative programs." She questions the assumption that "new is better" and that "exposure brings improvement." She goes on to say that: "Without descriptions of actual training activities, analyses of what specific needs they addressed, and more precise definitions of their effects, it is hard to interpret the findings in this evaluation."[3]

It is hard for Feiman to interpret the data because she has a different set of questions than those answered in the cited research. But a simple syllogism may make the data quite interpretable for others.

> Participation in training sessions is improvement.
> Twenty-seven percent participated in training.
> Twenty-seven percent are improved.

Perhaps this is a poor definition of improvement -- but in whose opinion.

I showed the data paragraph from the pilot teacher center evaluation to a graduate student and asked "What's it mean?" The student replied: "40 percent of the respondents got graduate credit for their time. It means taxpayers are getting ripped off to raise the credentials of teachers so they'll get higher salaries -- under the guise of teacher improvement."

This long discourse on the perception of flowers versus maggots is not aimed at simply making the point that different people perceive things differently. I take that to be a truism. The point is that this truism is regularly and consistently ignored in the design of evaluation studies. Laboring under the assumptions of positivistic social science, we pretend that there is some body of data out there that has only to be collected at which point we will know what it all means, whether or not it works, whether or not we're successful, whether or not

---

[3]Sharon Feiman, "Evaluating Teacher Centers," School Review (May, 1977), in press.

we're effective. Such data simply do not exist outside the context of a specific group of people with a particular perspective.

The implications for evaluation of this truism about differential perceptions are enormous. It means that instead of beginning with the traditional first question in evaluation textbooks -- "What are the program's goals?" -- we must begin by asking: "Whose goals for the program will be evaluated?" Instead of asking "What's it mean?" we must ask "Who is going to decide what it means?" Before exploring the implications of this perspective let me try to make the point a different way.

Hunting Bears

Zen Koans are one educational technique for helping the learner look at the world in a different way. Another technique that has emerged over time through journeys into and out of a variety of rich cultural traditions is the Sufi story. Sufi stories, particularly stories about the adventures and follies of the incomparable Mulla (Master) Nasrudin, are a means of communicating ancient wisdom. "Nasrudin is the classical figure devised by the dervishes partly for the purpose of halting for a moment situations in which certain states of mind are made clear.... Since Sufism is something which is lived as well as something which is perceived, a Nasrudin tale cannot in itself produce complete enlightenment. On the other hand, it bridges the gap between mundane life and a transmutation of consciousness in a manner which no other literary form yet produced has been able to attain."[4] What, then, can we learn from Nasrudin about evaluation of teacher centers? Let's try the parable about hunting bears.

> A king who enjoyed Nasrudin's company, and also liked to hunt, commanded him to accompany him on a bear hunt. Nasrudin was terrified.
>
> When Nasrudin returned to his village, someone asked him: "How did the hunt go?"

---

[4]Idries Shah, The Sufis (New York: 1964), p. 56.

"Marvelously!"

"How many bears did you see?"

"None."

"How could it have gone marvelously, then?"

"When you are hunting bears, and you are me, seeing no bears at all _is_ a marvelous experience."[5]

Permit me to translate this tale into modern language. The evaluation report on the king's hunt might read something like this:

EVALUATION OF THE BEAR PROJECT

This is a study undertaken for His Majesty's Ministry of the Interior, under the auspices of the Department of Natural Resources, for the Division of Parks, Section on Hunting, Office of Bears. This is a study of the relationship between the number of bears sighted on a hunt and the number of bears shot on a hunt. Our hypothesis is that there is a direct, linear relationship between the sighting of bears and killing of bears. The data was collected on a recent royal hunting expedition. The sample size is therefore somewhat small and generalizations cannot be made with confidence. In effect this is an exploratory case study, Campbell and Stanley Research Design No. 1.

The data support the hypothesis at the .001 level of significance. Indeed, the correlation is perfect. The number of bears sighted was zero, and the number of bears killed was zero. In no case was a bear killed without first being sighted. We therefore recommend that in future projects new Royal regulations be implemented requiring that bears first be sighted before they are killed.

Respectfully submitted,

The Incomparable Mulla Nasrudin

Royal Evaluator

---

[5] _Ibid._, p. 61.

While this evaluation report may be statistically somewhat less rigorous than the average evaluation study, it shares one major characteristic with almost all other reports of this genre. It is impossible to tell whether or not it answers anyone's question. Who decided that the outcome should be the number of bears killed? Perhaps the project staff simply uses the hunt as a format for getting royal (Federal) money to conduct field trips and the real outcome is a heightened sensitivity to nature? or a closer relationship between Nasrudin and the king? or a reduction of Nasrudin's fear of bears? or an increase in the king's power over Nasrudin? It may even be possible (likely!) that different characters in the situation have different objectives and would like different outcome measures. If so, it seems unlikely that all characters will be interested in the same evaluation data. Who will decide what it all means? For Nasrudin the data indicated a "marvelous" outcome. Other decision-makers might read the data differently.

Utilization of Evaluation Research

This emphasis on determining who the evaluation will serve as the first step in the evaluation process derives from a personal bias that evaluations ought to be useful -- they ought to have an impact on what people think, plan, and do. In 1973 this bias led me to get involved in a study of the factors that affect the utilization of evaluation research. We studied the utilization of twenty Federal health evaluations by interviewing the principle program person, evaluator, and project officer in each case.[6] In attempting to identify the factors that affect utilization of evaluation findings we asked respondents to comment on the relevance and importance of eleven factors extracted from the literature on utilization:

---

[6]The study is described in detail in Michael Patton, et.al., "In Search of Impact: An Analysis of the Utilization of Federal Health Evaluation Research," in Using Social Science Research in Policy-Making, Carol Weiss, ed. (New York, 1977), pp. 141-163.

> methodological quality, methodological appropriateness, timeliness, lateness of report, positive-negative findings, surprise of findings, central-peripheral program objectives evaluated, presence-absence of related studies, political factors, government-evaluator interactions, and resources available for the study.

Finally, we asked respondents to "pick out the single factor you feel had the greatest effect on how this study was used."

Two related factors emerged as important in our interviews: (1) a political considerations factor and (2) a factor we have called the personal factor. This latter factor was unexpected and its clear importance to our respondents has substantial implications for the utilization of evaluation research. None of the other specific literature factors about which we asked questions emerged as important with any consistency.

The personal factor emerged most dramatically in our interviews when, having asked respondents to comment on the importance of each of our eleven utilization factors, we asked them to identify the single factor that was most important in explaining the impact or lack of impact of that particular study. Time after time, the factor they identified was not on our list. Rather, they responded in terms of the importance of individual people.

For lack of a better term, we have simply called this new variable the personal factor. It is made up of equal parts of leadership, interest, enthusiasm, determination, commitment, aggressiveness, and caring. Where the personal factor emerges, evaluations have an impact; where it is absent, there is a marked absence of impact.

Comments from one study with an unusually high level of utilization are illustrative. The decisionmaker was asked why this particular study had so much impact. His answer was brief and to the point:

> Well, [the evaluation had an impact] because we designed the project with an evaluation component in it, so we were expected to use it and we did....Not just the fact that [evaluation] was built in, but the fact that we built it in on purpose. That is, the agency head and myself had broad responsibilities

> for this, wanted the evaluation study results and we expected to use them. Therefore they were used. That's my point. If someone else had built it in because they thought it was needed, and we didn't care, I'm sure the use of the study results would have been different.[7]

The evaluator, (an external agent selected through an open RFP process), completely agreed that:

> The principal reason [for utilization] was because the decisionmaker was the guy who requested the evaluation and who used its results. That is, the organizational distance between the policy maker and the evaluator was almost zero in this instance. That's the most important reason it had in impact.[8]

What emerges here is a picture of a decisionmaker who knew what information he wanted, an evaluator committed to answering the decisionmaker's question, and a decisionmaker committed to using that information. The result was a high level of utilization in making a decision contrary to the decisionmaker's initial personal hopes.

This point was made often in the interviews. One highly placed and highly experienced administrator from yet a different project offered the following advice at the end of a four hour interview:

> Win over the program people. Make sure you're hooked into the person who's going to make the decision in six months from the time you're doing the study, and make sure that he feels it's his study, that these are his ideas, and that it's focused on his values....[9]

The message here is in two parts. First, evaluators cannot conduct useful studies unless they know whose question they are answering; and secondly, program people have a right and an obligation to participate in the entire evaluative process from design to interpretation if results are to be useful. In some cases such participation means making demands that are unusual and beyond the experiences of most evaluators. In 1972 the teacher training program at the New School for Behavioral Studies in Education, University of North Dakota, was to be evaluated

---

[7] Ibid., p. 157.

[8] Ibid., p. 158.

[9] Ibid.

as part of a national Office of Education study. Dean Vito Perrone argued that the study, as designed, would be useless to the New School. He talked the Office of Education people into allowing him to spend the New School's portion of the evaluation money on a study locally designed and locally conducted. The subsequent evaluation was entirely staff designed and produced instruments and data that have become an integral part of the North Dakota program. The national study produced large volumes of numbers (with blanks entered on the lines for North Dakota), and as far as I can tell, was of no particular use to anyone.

Contemplate, then, the usefulness of evaluation research. Begin the evaluative process by asking: Who is this study for? Whose question will be answered? Begin by contemplating the usefulness of what to whom. And if you can't answer that initial question in terms of real people with real questions, then leave the evaluation and contemplate a Zen Koan; for example, the Koan on snow: "Where do all these lovely snowflakes fall, anyway? Do they fall in any particular place?"

PROGRESS VERSUS CHANGE

One of the purposes of evaluation research, for some the major purpose, is to tell us whether or not we're getting where we want to be. In the preceding pages I've suggested that different people involved in the same program may want to get to different places. Various people will interpret data about where we are in different ways. For many the issue centers on the question of improvement. Are things improving? Are teachers improving? Are children doing better? These may well be the central questions in an evaluation of teacher centers. Kathleen Devaney in a paper on Teacher Centers argues that assessment of program effectiveness "will have to be in terms of verifying individual teachers' improvement in classroom performance over time...."[10]

---

[10]Kathleen Devaney, "Surveying Teachers' Centers," Teachers' Centers Exchange, Occasional Paper No. 1, National Institute of Education, (April, 1977), p. 7; italics added.

Let me suggest that for teacher center evaluations it may be important to separate the issue of improvement from the related, but quite different, issue of impact or change. Improvement involves a judgment about whether or not something is better. Impact involves the more limited question of whether or not something is different. An observed difference may or may not constitute improvement depending on who is making the value judgment about whether or not a change is for better or worse. It is crucial throughout the evaluation process that empirical observations about program impact be kept separate from judgments about whether or not such impact constitutes improvement.

Suppose a teacher center conducts a series of workshops on the use of resources outside the classroom. As a result, a group of teachers increase the time students spend in the community by an average of three hours a week. The time spent outside the classroom led to an average reduction of one hour per week in time spent in class work doing both reading and arithmetic. Outside the classroom the students are exposed to a variety of activities and opportunities, but there is a net reduction in the time spent doing supervised reading and arithmetic. Clearly the teacher center has had an impact. Change has occurred. But have the teachers "improved"? The answer to that question depends on how much one values supervised reading and arithmetic compared to other stimulating activities.

A Sufi story illustrates quite nicely the problem of determining whether or not improvement has occurred. The setting is a teahouse, a Sufi term for a meeting place of dervishes. A monk enters and states:

> "My master taught me to spread the word that mankind will never be fulfilled until the man who has <u>not</u> been wronged is as indignant about a wrong as the man who actually <u>has</u> been wronged."

The assembly is momentarily impressed. Then Nasrudin speaks:

> "My master taught <u>me</u> that nobody at all should become indignant about anything until he is sure that what he thinks is a wrong is

in fact a wrong -- and not a blessing in disguise![11]

Questions of right and wrong, better or worse, are not simple empirical questions. To formulate evaluation questions solely in such terms can sabotage an evaluation from the beginning. What then can one do? In my judgment, the empirical question is not improvement but change. I suggest that we begin not with the question of whether or not teachers are "better," but whether or not they are different. Has the program been effective in changing teachers? Do they think differently? Can they do things now that they couldn't do before? Do they feel differently? Are different things occurring in teachers/ classrooms? These are empirical evaluation questions. Data from these evaluation questions can then be used to determine whether or not such changes and differences constitute progress or improvement.

This is not an esoteric, semantic distinction. Nor is it the beginning of a polemic on value-free social science. It is a practical suggestion for distinguishing between that which can be observed (by whatever methods) and that which cannot be observed. Failure to make that distinction can lead to serious misunderstandings and miscommunications throughout the evaluation process.

What, then, can be observed? There are essentially five categories of changes in teachers one might investigate:

1. Changes in teachers' feelings (i.e., morale, support, enthusiasm, isolation, etc.);
2. Changes in teachers' opinions (i.e., interpretations, attitudes, expectations, goals, etc.)
3. Changes in teachers' knowledge (i.e., facts, data, theories, models, etc.);
4. Changes in teachers' skills (i.e., things they can do); and
5. Changes in teachers' behavior (i.e., things teachers actually do do in teaching).[12]

---

[11] Shah, op. cit., pp. 58-9.

[12] These five categories correspond to the five elements of awareness identified by Elam Nunnally, Sherrod Miller, and Daniel Wackman, Alive and Aware.

A particular teacher center might focus on any or all of these types of changes. Programs typically vary considerably in degree of emphasis on different outcomes. A comprehensive evaluation might well look at all of these outcomes. For some educators "improvement" only occurs if teacher behavior changes; for others changes in teachers' feelings constitute sufficient improvement to justify a program. Many educators relate changes together in a sequence: knowledge changes (new information) then attitude change (new understandings) then skill change (new capacity) then behavior change (new actions). Where in this chain does real improvement begin? Kathleen Devaney quotes one educator to the effect that:

> The key to opening people to change is attitude, not information and skills. I can distill into ten pages what I think a teacher needs [in order] to teach reading. But before that, teachers must get a set of basic attitudes -- how language is viewed, how reading is viewed. Attitudes are the crucial foundation that makes it possible to assimilate new skills.

By carefully specifying the changes that staff want to observe in teachers as a result of teacher center activities, program staff can clarify and make explicit what they mean by "improvement." The next task is to determine what different levels of change indicate about the program. If fifty-five percent of the teachers in a workshop actually use a set of materials in their classroom, is that high usage or low usage? Does that mean the workshop was effective or ineffective? The workshop had an impact, but what level of impact is desirable? What level is acceptable? And what level spells trouble? These issues should be resolved before the data is collected to permit the discussion about what constitutes improvement to take place in an atmosphere that is not charged with defensiveness, rationalization, and justification.

Suppose you are collecting data on frequency of individual visits as Feiman did.[14] Her data can be grouped as follows:

---

[13] Op. Cit., p. 7.

[14] Sharon Feiman, Teacher Curriculum Work Center: A Descriptive Study, North Dakota Study Group on Evaluation Monograph, University of North Dakota (Grand Forks, N.D., 1975), pp. 19-21.

Number of Visits Made by a Teacher to the Center	Number of Visits	% of Total Visitors
1 - 2	185	80.4%
3 or more	45	19.6%

Note: Data are for visits between January 10 and February 28.

I don't know how staff reacted to these findings. My guess is that once these data are collected the program staff might look at them and say, "Oh, yes, that's about what we've anticipated. Plus the data don't include people who come to regular workshops and special classes. Then, too, 'since only 23 teachers noted on the background forms that they first visited the Center during the period of observation, it is not likely that most of the people who came once were first-time visitors.' And the observation time was really too short. Plus January and February and bad months, you know, everyone is depressed with winter, and...."

Soon it becomes apparent that either the data don't tell us much, at least not without other data, or that staff aren't prepared to deal with what the data do suggest. This is not at all unusual as a post-evaluation scenario.

Now let us try another scenario. At the outset of the evaluation study the program staff discuss their notions of what their task is and how teacher change occurs. The staff decide that the kind of impact they want to have cannot occur in one or two visits to the teacher center. "If teachers don't return after one or two visits we must be doing something wrong." The period of time in question is a full twelve month period. Before the data are collected the staff complete the following table:

Interpretation of data	Percentage of teachers who have contact with the Center three or more times
We're doing an outstanding job of engaging teachers at this level	51 - 75%
We're doing an adequate job of engaging teachers at this level	26 - 50%
We're doing a poor job of engaging teachers at this level	0 - 25%

A record keeping system must then be established that staff agree to and believe in so that the data have credibility. The teacher center staff have committed themselves to actively engaging teachers on a multiple contact basis. The data will provide clear feedback about the effectiveness of the program. The key point is that if staff are unwilling and/or unable to interpret data and set expectancy levels before the evaluation there is no reason to believe they can do so after the evaluation. In addition, going through this process ahead of time alerts participants to additional data they need in order to make sense of the evaluation; clearly one table on frequency of visits is only a starting place. Finally, involving staff in such a process helps clarify the evaluation criteria that are being used and, if staff are involved in establishing these criteria themselves, the evaluative process will increase their commitment to use the data for program improvement.

If multiple contacts with teachers is a key issue for center staff they would probably want to conduct follow-up interviews with teachers who didn't return and teachers who did return. What was the difference? There are many possibilities for extending this example. The point is that the evaluation question must be carefully formulated in accordinace with the basic goals of the teacher center; staff are involved in establishing explicit criteria for interpreting the data; and staff commit themselves to taking the process seriously.

Many of the most serious conflicts in evaluation research are rooted in the failure to clearly specify criteria in advance of data collection. This can lead both to collection of the wrong data and/or intense disagreement about the standards for interpreting data that have already been collected. Without explicit criteria data can be interpreted to mean almost anything about the program -- and about the quality of teacher center staff. Witness, for example, Master Ōbaku's famous evaluation of the staff in eighteenth century Chinese teacher centers:

> Master Ōbaku said, "You are all leftover eaters! If you walk around the world and search for truth in such a manner, what achievement can you expect? Do you know that there are no more Zen masters in China?"
>
> Then a monk stepped out and said, "Aren't there those who walk around earnestly instructing the masses? What of them?"
>
> Ōbaku said, "I did not say there is no Zen anymore, only that there are no great masters."

The Outcomes of Individualization

Feiman reports that most teacher center programs advocate individualization in that "teachers must be allowed to begin at their own beginnings, draw on personal strengths, and learn at their own pace."[15] This creates special problems in evaluating the outcomes of teacher center experiences, particularly problems in interpreting varying levels of change to make interpretations about improvement. Measuring improvement for evaluation purposes turns out to be a very tricky problem. The first problem is determining the level at which learners enter the situation. Then, for any given knowledge area or skill there are upper limits of competency. One can't show improvement indifinitely. It is well-established that learning occurs most rapidly and in greatest amount for a person who is just beginning to study a new subject. The reason is simply that there is more to

---

[15]Feiman (1977), Op. Cit., p. 16.

learn at the beginning. After a time, percentage gains decline. Persons who have never driven a car can improve their skills 100%; persons who have driven for years may only be able to show a small percentage gain after further instruction. How can the large percentage gains in a program of novices be compared to the small percentage gains in a program of experts? And teacher centers include both!

An individualized teacher center program requires an individualized set of evaluation change criteria that take into consideration where a teacher begins, what the teacher wants and needs, and what changes occur after the experience. This may mean a careful system of descriptive records for teacher center participants (or a sample thereof) not unlike the record-keeping systems developed by Pat Careny[16] for observing changes in students over time. Such a system permits the recording of some types of information about each teacher but also permits the collection of systematic data on the individual developement of each teacher. Such records serve program advisors for both planning and evaluation purposes.

The point here is that an evaluation system is worthless if it collects data about the wrong things. Observing changes in a highly individualized program means monitoring and describing different changes for different individuals and then looking for overall patterns of teacher change and learner activity in the program.

One evaluation report on a teacher workshop program tried to solve the problem by simply asking teachers how many felt they had changed as a result of their workshop experience. Seventy-six percent said they had changed. But no data was collected on how they changed or what activities led to the change. In terms of actual program planning and improvement the data weren't very useful.

An oft-repeated Sufi story illustrates the importance of observing and collecting data on the right thing.

---

[16]Pat Careny, A Phenomenological Approach to the Study of Human Phenomena. North Dakota Study Group on Evaluation, University of North Dakota, (Grand Forks, 1975.

Nasrudin used to take his donkey across a frontier every day, with the panniers loaded with straw. Since he admitted to being a smuggler when he trudged home every night, the frontier guards searched him again and again. They searched his person, sifted the straw, steeped it in water, even burned it from time to time. Meanwhile he was becoming visibly more and more prosperous.

Then he retired and went to live in another country. Here one of the customs officers met him, years later.

"You can tell me now, Nasrudin," he said. "Whatever was it that you were smuggling, when we could never catch you out?"

"Donkeys," said Nasrudin.

The One True Way

Enlightenment is the goal in Buddhism. Koans and Sufi stories are only two of many techniques for assisting the learner along the path toward Enlightenment, "the shattering of the Great Doubt..., the welling up of a flood of exaltation. 'If you take up one Koan,' Master Hakuin said, 'and investigate it unceasingly, your mind will die and your will will be destroyed. It is as though a vast, empty abyss lay before you, with no place to set your hands and feet. You face death and your bosom feels as though it were fire. Then suddenly you are one with the Koan, and body and mind are cast off.... You must push forward relentlessly and with the help of this complete concentration you will penetrate without fail to the basic source of your own nature."[17]

The Koans were a method for attaining Enlightenment, but they were also an evaluation tool for detecting Enlightenment. There are correct responses and incorrect responses to the Master's Koanic question, "What is the sound of one hand clapping?" There are close to two thousand recorded Koans -- complete sequences of questions and enlightened responses.[18] Koans are supposed to be passed on in secret from Master to disciple, "transmitted only by word of mouth

---

[17]Hoffman, Op. Cit., pp. 22-23.

[18]Ibid., p. 39.

to the student in the master's room.[19] The pupil vows to keep the Koans a secret and to transmit them only to chosen disciples after he has become a Zen Master.

Koans were first published under a pseudonym by a rebel Zen master who had become disenchanted with his fellow masters. The publication created an enormous scandal because once the Koans were made public anyone could supposedly memorize the correct responses. Therefore, anyone could become a Zen master. How, then, was one to tell the truly enlightened from the fakes? Meanwhile, the rebel master maintained that monks were already mindlessly repeating Koans without being truly enlightened.[20]

The Koans themselves emerged out of a debate over whether enlightenment came from passive meditation or a more active searching with the mind. Other Zen monastic retreats disagree over whether Enlightenment comes slowly over time or quickly, all-at-once. Of such issues are sects created, each teaching the One True Way.

Western social scientists would scoff at such religious absolutism and, under the enlightenment provided by cultural relativism, would explain in great depth why each culture deceives itself with its own version of Truth. These same social scientists will then design an educational experiment to determine the one best way to teach students that there is no one best way to attain Enlightenment.

When I finished reading Sharon Feiman's excellent paper on the three models of teacher centers[21] I immediately had a vision of a researcher writing a proposal which would be enthusiastically funded by the National Institute of Education to determine which of the three models was most effective in changing teachers. This despite the fact that every large scale experiment of this kind (witness the

---

[19]I. Miura and R.F. Sasaki, The Zen Koan (Kyoto: First Zen Institute of America, 1965), p. 28.

[20]Hoffman, Op. Cit., pp. 39-44.

[21]Feiman (1977), Op. Cit.

National Follow Through Evaluation) has found that the within-group variation is greater than the between-group variation, i.e., the "experiment" fails to identify a single best "treatment."

But the search goes on. Educational programs are compared to each other using standardized criteria despite the fact that different programs are trying to do different things. Such evaluation comparisons are particularly insidious because they subtly divert program staff from their original goals. Let me illustrate the problem using the Feiman paper.

Feiman describes three types of teacher centers: "behavioral" centers, "humanistic" centers, and "developmental" centers. I apologize for the injustice done to her paper by the table below, but this table represents a simplified summary of my understanding of the differences among the three models.

Type of Center	Primary Process of Affecting Teachers	Primary Outcomes of the Process
1. Behavioral Centers	Curriculum specialists directly and formally instruct administrators and teachers	Adoption of comprehensive curriculum systems, methods and packages by teachers
2. Humanistic Centers	Informal, non-directed teacher exploration; "teachers select their own treatment."	Teacher feel supported and important; pick up concrete and practical ideas and materials for immediate use in their classrooms.
3. Developmental Centers	Advisors establish warm, interpersonal, and directive relationship with teachers working with them over time.	Teachers' thinking about wh they do and why they do it is changed over time; teache personal developmen

Regardless of whether or not one agrees with Feiman's categories, it is clear that, at least to some extent, different teacher centers are trying to accomplish different outcomes. The three models cannot be compared to determine which one is best, or most effective, because they are trying to do different things. Evaluation can help determine whether or not each of the outcomes have been attained for each specific program, but evaluation cannot determine which

outcome is most desirable to attain or which outcome is "best."

Moreover, it is important to distinguish clearly between the outcomes desired for a program and the process used for attaining those outcomes. Feiman suggests that there are three models of teacher centers actually in operation each characterized by a specific process linked to a specific set of desired outcomes. But theoretically there are nine models, one model for each combination of process and outcome. In addition there are a nearly endless variety of mixes with some teacher centers undoubtedly using all three processes.

Another Sufi story illustrates the importance of understanding clearly the difference between the processes for attaining a goal and the actual goal itself.

> The incomparable Mulla Nasrudin is visited by a would-be disciple. The man, after many vicissitudes, arrives at the hut on the mountain where the Mulla is sitting. Knowing that every single action of the illuminated Sufi is meaningful, the newcomer asks Nasrudin why he is blowing on his hands. "To warm myself in the cold, of course."
>
> Shortly afterward, Nasrudin pours out two bowls of soup, and blows on his own. "Why are you doing that, Master?" asks the disciple. "To cool it, of course," says the teacher.
>
> At this point the disciple leaves Nasrudin, unable to trust any longer a man who uses the same process to arrive at different results -- heat and cold.

A comprehensive evaluation will gather information about both process and outcomes. Indeed, the processes identified by a local teacher center staff as those processes they desire to employ in their work constitute a set of program implementation goals. Evaluators refer to the study of factors affecting program implementation as "process evaluation." Feiman's study of the Teacher Curriculum Work Center[22] is a process evaluation. She studied what teachers did at the Center, i.e., the processes whereby they obtain support, ideas, and materials; she did not study whether or not teachers actually changed what they did in the classroom. Such process studies can be extremely valuable for they

---

[22]Feiman (1975), Op. Cit.

provide staff with information about whether or not the program is being implemented as desired.

Just as it is inappropriate to compare teacher centers that aim at different outcomes, it is also inappropriate to compare the teacher center approach to educational in-service with other school-based or University-based approaches to teacher in-service. To do so is to once again search for The One True Way in a world where different approaches work differently for different people. Each approach must be empirically evaluated in its own terms; the decision about whether or not the goals of a particular approach are desirable is a political question.

From whence cometh this evaluation preoccupation with identifying the <u>one</u> best approach to a problem? From the earliest days of my social science training I was taught that the aim of social science was to find that one set of variables that would allow us to perfectly predict human behavior. I was taught that the problem of perfectly predicting human behavior was largely a measurement problem. If we only had complete data on a person or group of people we could predict with near certainty the behaviors, attitudes, and inner workings of that person or group. The equations never quite worked out but that was because of "measurement error," problems with "variable specification," and, of course, the unfortunate fact that there will always be a few deviants around to screw up our predictions.

This positivistic perspective has had an important effect on governmental policy-making and on research funding, especially evaluation and policy research. Legislators would like to find the <u>one</u> solution, so they mandate research to tell them the answer -- the one answer -- to education, health, welfare, human relations, bureaucracy, ... Which educational model is best? What's the one best way to organize government? Which one program approach is most effective in reducing crime and rehabilitating criminals?

The truth is that there is no one best program approach across the board for the same reason that there is no single factor or set of factors that can solve

the mystery of human behavior, no one answer to that most fundamental philosophical question: Why do people do what they do? (And that most fundamental governmental question: How do we get people to do what we want them to do?)

The reason is that individual people are different from each other in unique ways -- and those unique individual characteristics make a difference in what people do and how they respond to programs. From the structuralist perspective that dominates social science, individual people do not make a difference. We learn in introductory sociology that the major characteristic of modern society as a rational system is the interchangeability of people in positions. It doesn't matter who is President; the position is determined by large-scale, long-term, socio-political-economic factors that are not subject to personal manipulation. It doesn't matter who the bureaucrat is in the welfare office; it matters only that someone, anyone fill the position. It doesn't matter who runs an educational program, the trick is to structure the program effectively.

But it's just not so. The prediction equations for human behavior don't work perfectly because human behavior is not perfectly predictable. After the last of background variables, demographic factors, and structural influences have been measured and entered into the equations, the accuracy of the predictions are downright puny. Freedom of individual action, personal predilections and choice, unique circumstances, and the fundamental individuality of human beings win the day.

In educational research we keep discovering that what makes the difference is the enthusiasm and dedication of individual teachers, not some theoretical model of instruction. In therapy and counseling what matters is the sincerity and caring of one individual trying to help another, not the psychotherapeutic method employed. In program after program I find that the abilities, personalities, and dedication of the people who run and staff programs make the major difference. In effect, the personal factor that I described in the first section of this paper as the

key factor in determining the utilization of evaluation research demonstrates the same insight. Individual people and individual circumstances make the major difference in what happens.

This means that instead of searching for that single model of education, health, or welfare that will work everywhere, legislators and government administrators must find ways to stimulate innovation and adaptation of programs to fit unique local needs and circumstances. Evaluation of such programs must also be based on the unique characteristics and goals of local programs, not evaluation by standardized criteria applied across the board.

The great threat from Big Government (or behavioral science in the service of Big Government) is not that it will ever perfectly control or manipulate us. The great danger is that we must use an ever greater quantum of our personal energy, creativity, and freedom of action finding ways to subvert standardized prescriptions for behavior (and evaluation) that are usually irrelevant and often destructive.

The search for solutions will go on to be sure, both in science and in education. But the search ought not be for the One Solution. Rather, we must search for situational solutions -- What is best for this program at this time in these circumstances? How can we help staff accomplish their goals? The challenge is to identify and implement alternatives not universal prescriptions that treat people as if they're all the same. This places the responsibility for program evaluation clearly at the local level.

This does not mean that there is no responsibility for evaluation at the national level. Quite the contrary, the Federal responsibility is to make available resources and incentives for the conduct of local program evaluations. Each local program should describe and evaluate both processes (what the program is doing?) and outcomes (what changes occur in teachers and classrooms?). These local evaluations will then provide rich case study data that can be content

analyzed at the national level to look for patterns, trends, and variations in effectiveness under different conditions. Therein will be found information for program planning and development.

Of course, this is not the easy way to carry out evaluations. It is hard on local staff because they must clearly and explicitly state objectives for both program processes and outcomes; and they must develop methods of describing, observing, and measuring those processes and outcomes. It is hard on Office of Education and N.I.E. staff because they cannot apply some handy yardstick to all programs and thereby discover which is best. It is hard on evaluators because they must interact intensively with local staff to determine what information is appropriate in a particular situation. But if teacher centers are evaluated as other national educational programs have been evaluated (e.g., Headstart, Follow Through, National Assessment), we will continue to be like Nasrudin looking for his lost key.

> On one occasion a neighbor found Nasrudin down on his knees under a street lamp looking for something.
>
> "What have you lost, Mulla?"
>
> "My key," said Nasrudin.
>
> After a few minutes of searching, the other man said, "Where did you drop it?"
>
> "In that dark pasture."
>
> "Then why, for heaven's sake, are you looking here?"
>
> "Because there is more light here."

On Causes and Effects

The reason for separating program processes from program outcomes is so that they can be related to each other. The sequence of questions goes like this:

1. Are we implementing the program the way we want to (process evaluation)?
2. Are teachers changing/developing the way we want them to (outcomes evaluation)?

3. Is what we're doing in the program (the processes) related to the changes we see in teachers (the outcomes)?

This last question takes us into the arena of causality. Do the processes/activities of the teacher center cause/affect the behaviors/attitudes/skills/knowledge/feelings of teachers? One need know very little about research to know that it is impossible to establish causality in any final sense when dealing with the complexities of real programs where treatments are never pure, single, and uncontaminated.

It is easy to become frustrated with the difficulty of establishing the relationship between program activities and program outcomes. We can't answer such questions definitively. No amount of scientific sophistication will provide definitive answers to causal questions for the same reason that there is no One right approach to changing human behavior. But that is no reason not to ask the question. We cannot provide definitive answers but we can arrive at some reasonable estimation of the likelihood that particular activities have had an effect.

This point is important for it tells us something about what we can expect from evaluation research. Evaluation data are never clear-cut and absolute. Studies are always flawed in some way. There are always questions of reliability and validity. Error-free instruments do not exist and cannot exist in the measurement of complex human social/behavioral/psychological phenomena. Of what good is all this then?

Evaluation research is only of use if you believe that some systematic information is better than no systematic information. Evaluation research only has meaning if you believe that a rough idea of the relationship between program activities and outcomes is better than reliance entirely upon hope and good intentions. Evaluation research does not provide final answers, but it can provide direction.

Earlier I referred to a study aimed at determining the factors that affect

the utilization of evaluation research.[23] The first task in that study was to describe how evaluations actually were used. What we found was considerably more complex and less dismal than our original impressions led us to expect. Evaluation research is used but not in the ways we had anticipated. None of the impacts described was of the type where new findings from an evaluation led directly and immediately to the making of major, concrete program decisions. The more typical impact was one where the evaluation findings provided additional pieces of information in the difficult puzzle of program action, thereby permitting some reduction in the uncertainty within which any decisionmaker inevitably operates.

This reduction of uncertainty emerged as highly important to decisionmakers. In some cases it simply made them more confident and determined. On the other hand, where the need for change was indicated, an evaluation study could help speed up the process of change or provide an impetus for finally getting things rolling.

Reducing uncertainty, speeding things up, and getting things finally started are real impacts -- not revolutionary, organization-shaking impacts -- but important impacts in the opinion of the people we interviewed. One administrator summarized this view both on the specific evaluation in question and about evaluation in general:

> Myself I have a favorable view toward evaluating. If nothing else it precipitates activity many times that could not be precipitated without someone taking a hard look at an organization. It did precipitate activity in [this program]. Some of the findings were not positive. Some of it was negative. At least something occurred that wouldn't have occurred if the evaluation hadn't taken place.[24]

It is crucial to set realistic expectations at the beginning of the evaluative process about what can be expected and how data will be used. Everyone involved should clearly understand that evaluation findings can provide important information

---

[23]Patton, et.al., Op. Cit.

[24]Ibid., p. 147.

for program improvement, but findings must be interpreted in the full context of the particular circumstances under which a program operates. Evaluation results from a single study ought not be the sole basis on which major program decisions are made.

It is important, then, to ask causal questions, even though evaluation data can only provide an approximate picture of what is really happening. It is important to ask causal questions. It is also important to interpret the results with prudence and care.

> One day an old man approached Zen master Hyakujō. The old man said, "I am not a human being. In ancient times I lived on this mountain. A student of the Way asked me if the enlightened were still affected by causality. I replied saying that they were not affected. Becuase of that I was degraded to lead the life of a wild fox for five hundred years. I now request you to answer one thing for me. Are the enlightened still affected by causality?"
>
> Master Hyakujō replied, "They are not deluded by causality."
>
> At that the old man was enlightened.
>
> Master Hyakujō clapped his hands, laughed, and said, "I only thought that the barbarian's beard was red, I never realized it was a red-bearded barbarian."[25]

QUESTIONS FOR TEACHER CENTER EVALUATION: AN OVERVIEW FROM AN IRON COW

I began this paper with the basic tenet of the Koan educational method that the question is as important as the answer. This paper has centered on evaluation questions for teacher centers. In this final section I shall attempt to summarize and order those questions.

1. Who is the study for? Whose question will be answered? Who will determine what it all means? Begin by contemplating the usefulness of what to whom.

    a. Is the evaluation for local program staff to use in improving the program?

    b. Is the evaluation for state or federal officials to use in making global program decisions?

---

[25]Hoffman, Op. Cit., p. 138; adapted.

c. What expectations are there about how the evaluation data will be used, by whom, to make what decisions, on what issues?

2. What do we want to affect in teachers? How do we want teachers to be different after the program?

   a. What changes, if any, do we expect in teachers' feelings?

   b. What changes, if any, do we expect in teachers' opinions?

   c. What changes, if any, do we expect in teachers' knowledge?

   d. What changes, if any, do we expect in teachers' skills?

   e. What changes, if any, do we expect in teachers' behavior?

3. How will we observe, describe, and/or measure the degree to which teachers are different after the program?

4. What are our criteria for success? and failure? At what level and to what degree do we expect to affect teachers (for each specific outcome desired)?

   a. At what level are we doing an outstanding job?

   b. At what level are we doing an adequate job?

   c. At what level are we doing a poor job?

5. What program activities and processes do we expect to engage in to affect teachers?

   a. Do we expect to use formal instruction by curriculum specialists to affect teachers?

   b. Do we expect to use informal, non-directed teacher exploration to affect teachers?

   c. Do we expect to use advisors in establishing warm, interpersonal, and directive relationships with teachers?

6. How will we observe, describe, and/or measure the degree to which we actually implement program processes and activities as planned?

7. What are our criteria for success? and failure? At what level and to what degree do we expect to implement program processes and activities?

   a. At what level are we doing an outstanding job?

   b. At what level are we doing an adequate job?

   c. At what level are we doing a poor job?

8. What is the relationship between what we're doing in the program (the processes) and the changes observed in teachers (the outcomes)? Is there a relationship?

   a. How much confidence do we have in the strength of that relationship?

9. What's it all mean? What do we do? What have we learned about the program that gives us a direction for action?

10. The Koan of An Iron Cow

    Master Fuketsu said, "Buddha-mind is just like an iron cow; if there is movement--there is no progress; if there is standstill--there is stagnation. Well, this 'no-movement-no-standstill,' should one be mindful of it? Should one be unmindful of it?

    Answer: It's like a stone mill.

    Master: Why is it so?

    Answer: It doesn't move a bit. Along the hedge chasing a butterfly, by the water's side toying with a frog.

    Master: This 'no-movement-no-standstill,' should one be mindful of it? Should one be unmindful of it? If I keep after you in this way, how will you answer?

    Answer: The pupil slaps his master once.[26]

---

[26]Hoffman, Op. Cit., pp. 134-5.